คอมพิวเตอร์ กับ ภาษา

ยืน ภู่วรวรรณ

ความมหัศจรรย์ของสมองมนุษย์ในความเข้าใจภาษาเป็นสิ่งหนึ่งที่นักคอมพิวเตอร์ในยุคปัจจุบันกำลังให้ความสนใจ ทั่วโลกมีภาษาที่ใช้พูดกันมากกว่า 4,000 ภาษา ในบรรดาภาษาต่าง ๆ เหล่านี้ บางภาษาก็ใช้อยู่ในกลุ่มชนหมู่น้อย และไม่มีตัวเขียนเป็นของตนเอง อย่างไรก็ตาม ความเข้าใจภาษาของมนุษย์ก็เป็นสิ่งที่ลึกซึ้งเกินกว่าที่เครื่องจักรในยุคสมัยปัจจุบันจะทำได้

หากจะพิจารณาถึงความก้าวหน้าทางวิทยาการของมนุษย์ ปัจจุบันเรามีเครื่องคอมพิวเตอร์ที่มีความจำที่สามารถจดจำข้อมูลได้เป็นพัน ๆ ล้านตัวอักษร และมีขีดความสามารถในการคำนวณได้รวดเร็วหลายสิบล้านคำสั่งต่อวินาที ขีดความสามารถนี้มีแนวโน้มที่ก้าวหน้าขึ้นอย่างรวดเร็ว และเพิ่มขึ้นด้วยอัตราก้าวหน้าแบบเอ็กซ์โพเนนเชียล ด้วยเหตุนี้ความใฝ่ฝันของมนุษย์ที่จะให้เครื่องจักรทำความเข้าใจในภาษามนุษย์ จึงเป็นสิ่งหนึ่งที่จัดอยู่ในอันดับของการเร่งการพัฒนาเทคโนโลยีในยุคสมัยที่จะมีมาต่อไป ในบทความนี้จะกล่าวถึงขอบข่ายของการประมวลผลภาษาธรรมชาติที่จะเข้ามามีบทบาท โดยเน้นถึงหลักการพื้นฐานและความเป็นไปได้เพื่อเป็นแนวทางสำหรับผู้สนใจต่อไป

เริ่มจากการประมวลผลข้อความ เอกสาร (Text Processing)

การประยุกต์ใช้งานคอมพิวเตอร์ในเรื่องการประมวลผลภาษาธรรมชาติในยุคแรก ที่ประสบผลสำเร็จเห็นจะได้แก่การใช้ในการจัดการเอกสาร ข้อความ ปัจจุบันเรามีเอดิเตอร์ที่ทำการจัดการเอกสารด้วยรูปแบบและหลักการต่าง ๆ มากมาย การจัดการเอกสารเหล่านี้มีประโยชน์มากต่อการทำงานในเกือบทุกวงการ

จุดมุ่งหมายของการประมวลผลเอกสารนั้นทำให้โปรแกรมที่พัฒนามาดีขึ้นเป็นลำดับจนในที่สุดกลายเป็นเวิร์ดโปรเซสเซอร์ที่มีขีดความสามารถต่าง ๆ มากมาย เช่น เวิร์ดสตาร์ เวอร์ชัน 5 สามารถจัดระบบเอกสารได้เพิ่มขึ้นกว่าเดิมมาก มีพร้อมทั้งพจนานุกรมและระบบชวเลข

ความต้องการเวิร์ดโปรเซสซิ่งนั้น ผู้ใช้มีความต้องการที่สำคัญหลายอย่าง พอสรุปเป็นสังเขปได้ดังนี้

- ระบบควรมีขีดความสามารถสูงพอที่จะให้ผู้ใช้ทำการเปลี่ยนแปลงรูปแบบเอกสารได้ตามใจชอบ โดยทั่วไปจะเสมือนมีเครื่องมือพิเศษให้ เช่น ดินสอ ไม้บรรทัด กรรไกร และเทปกาว ฯลฯ

- ระบบควรมีความเร็วในการตอบสนองผู้ใช้ได้รวดเร็วพอเพียง เพราะการดำเนินการของผู้ใช้จะกระทำเสมือนด้วยเวลาจริง

- ระบบการเชื่อมโยงกับผู้ใช้จะต้องง่าย ใช้คีย์บอร์ดหรืออุปกรณ์ควบคุมได้โดยไม่ยาก ปัจจุบันมีการใช้เทคนิคหลายอย่างประกอบกัน เช่น ใช้ระบบเมนู อาจเป็นแบบพูลดาวน์เมนู มีป๊อบอัพวินโดว์ เป็นต้น

- การค้นหาข้อมูลในระดับทั้งหมด ควรจะทำได้รวดเร็วและสะดวก นอกจากนี้บางโปรแกรมยังเน้นการแยกคีย์เวิร์ด

และการค้นหาคีย์เวิร์ด รวมทั้งการทำสารบัญและดัชนีด้วย

- ในขณะทำงาน ผู้ใช้สามารถตรวจสอบเอกสารและแก้ไขได้อย่างสะดวก อย่างไรก็ตาม ด้วยขีดความเร็วของคอมพิวเตอร์ในปัจจุบัน ทำให้มีการเพิ่มขีดความสามารถพิเศษได้หลายอย่าง เช่น การแสดงผลแบบเหมือนบนกระดาษ

- ฯลฯ

เวิร์ดโปรเซสเซอร์สมัยใหม่ใส่หลักการ NLP ลงไปพอสมควร

หากลองพิจารณาขีดความสามารถของเวิร์ดโปรเซสเซอร์ตัวใดตัวหนึ่งที่เป็นเวอร์ชันล่าสุด เช่น เวิร์ดสตาร์เวอร์ชัน 5 เราจะเห็นหลักการของ NLP แฝงอยู่มากมาย เช่น การตรวจสอบตัวสะกด การหาความหมายคำ การให้พจนานุกรมคำพ้อง การใช้ชวเลข การคำนวณในเอกสาร ฯลฯ อย่างไรก็ตามมีผู้สร้างโปรแกรมสนับสนุนด้วยหลักการของ NLP มากขึ้น เช่น การตรวจสอบไวยากรณ์และสไตล์ของประโยค เป็นต้น

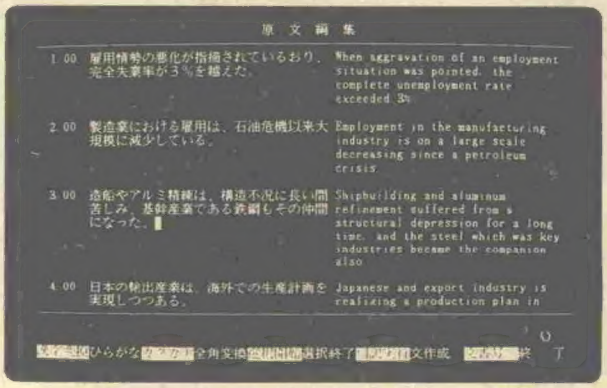
การจัดวางโครงร่างของเวิร์ดโปรเซสเซอร์

มีผู้ศึกษาวิจัยโครงสร้างของเอกสารที่จะใช้ในการประมวลผลกันมากพอควร การศึกษาพยายามหารูปแบบของโครงสร้างข้อมูลที่เหมาะสมต่อการประมวลผล เช่น เอกสารที่จะประมวลผล แบ่งย่อยเป็น บทที่ ตอน ย่อหน้า และประโยค ดังนั้นจึงมีการจัดโครงสร้างเป็นทรี เพื่อให้มีการขยายเอกสารได้และสามารถประมวลผลได้ดีเช่นกัน

รูปที่ 2 เป็นตัวอย่างของโครงสร้างเอกสารที่มองแบบ NLP โดยแบ่งเอกสารเป็นโครงสร้างลำดับชั้น การจัดโครงสร้างของข้อมูลนี้เน้นเพื่อให้ประมวลผลในเรื่องต่าง ๆ ได้ง่ายขึ้น

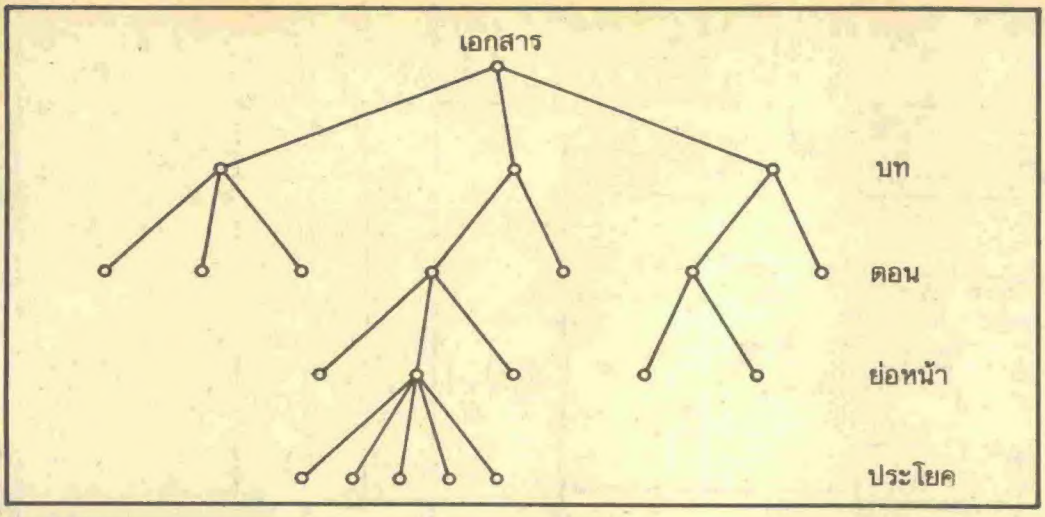
นอกจากนี้ มีการจัดเอกสารในรูปแบบอื่น ๆ อีกได้ เช่น ในโปรแกรม "hypertext" มีการจัดรูปแบบของข้อมูลเป็นบล็อกที่เชื่อมโยงกัน โดยแยกบล็อกเป็น 3 กลุ่ม คือ กลุ่มที่เป็น footnote กลุ่มคำอธิบายประกอบ (annotation) และ

ก. ระบบแปลภาษาของบริษัทโอกิ ที่ชื่อว่า PENSEE

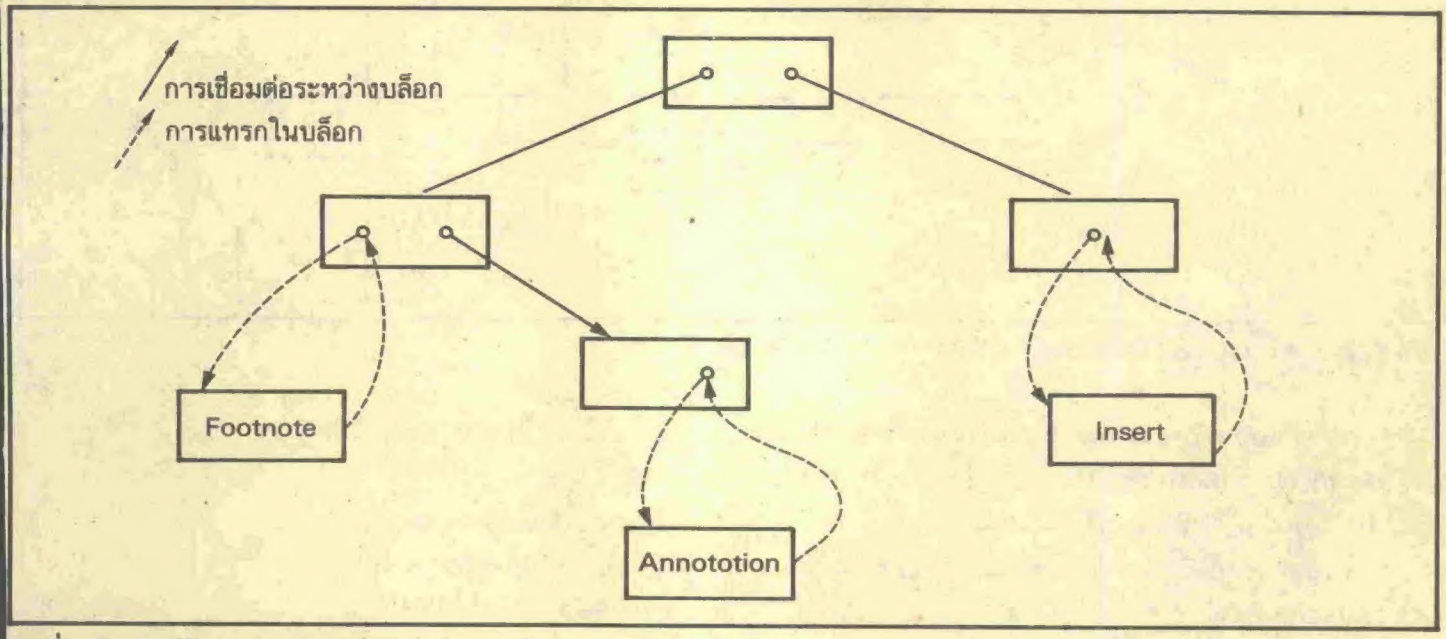


ข. ตัวอย่างการแปลจากภาษาญี่ปุ่นเป็นภาษาอังกฤษ

รูปที่ 1
การแปลภาษาด้วยคอมพิวเตอร์เป็นตัวอย่างของการประยุกต์ NLP งานแปลภาษาด้วยคอมพิวเตอร์เป็นงานที่มีผู้วิจัยกันหลายแห่งทั่วโลก



รูปที่ 2
โครงสร้างการจัดเอกสารระบบทรี



รูปที่ 3 โครงสร้างเอกสารในเวิร์ดโปรเซสเซอร์

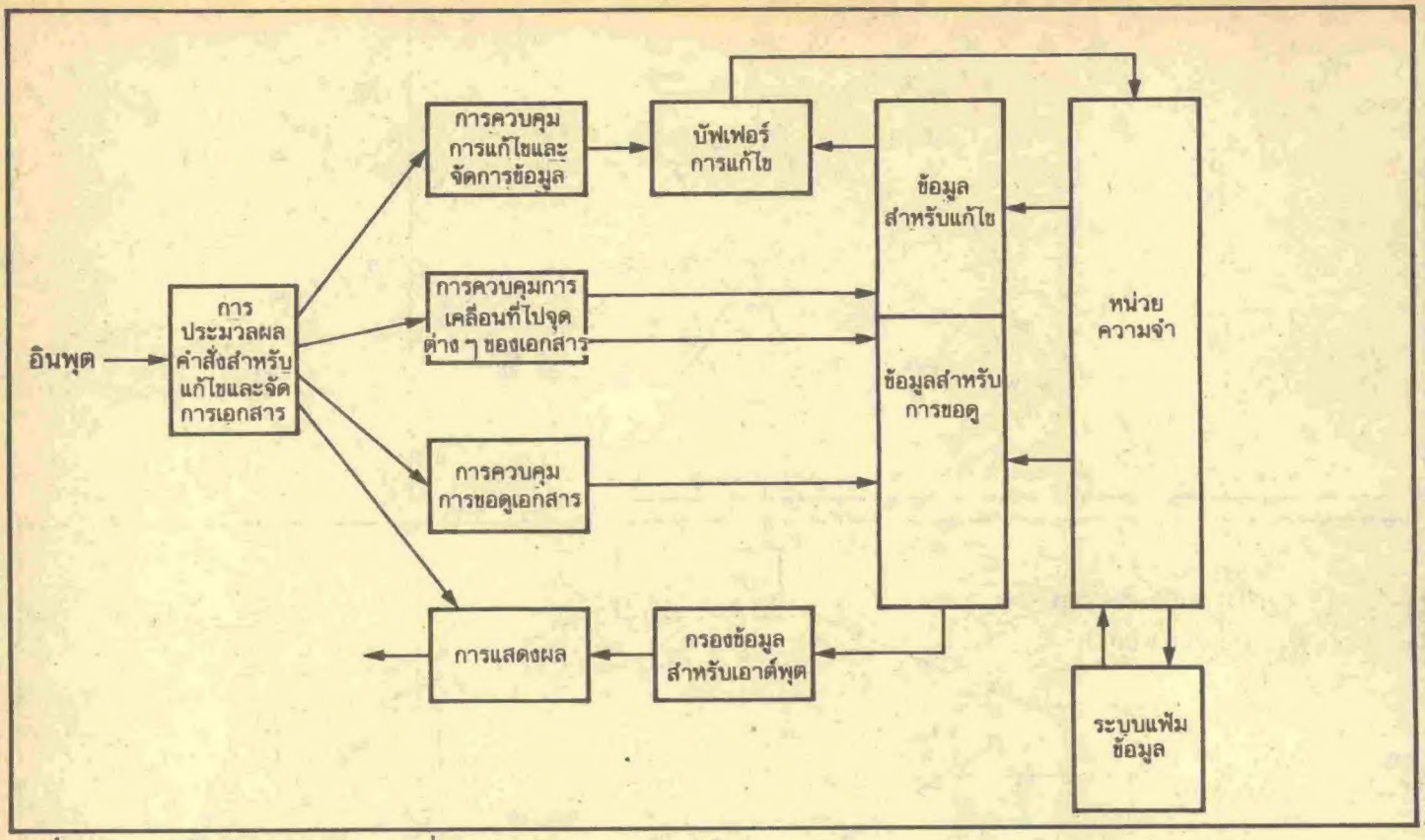
กลุ่มข้อความ การจัดการนี้จะมองข้อมูลเป็นบล็อก เมื่อมีการเพิ่มเติมข้อมูลก็จะแทรกเข้าในบล็อก โครงสร้างของการจัดการข้อมูล แสดงดังรูปที่ 3

การจัดการเอกสารแบบออนไลน์

การจัดเอกสารมีระบบที่จะกระทำได้ 2 แบบ คือ ออนไลน์กับออฟไลน์ ระบบออนไลน์จะกระทำตามแบบเวลาจริง เพื่อจะจัดการข้อมูลเอกสารโดยตรง เช่น การขอแก้ไขโดยการเลื่อนเคอร์เซอร์ไปยังตำแหน่งที่ต้องการและแก้ไข การขอดูเอกสาร รูปที่ 4 เป็นระบบการจัดการแก้ไขเอกสาร ซึ่งจัดเป็นโครงสร้างของระบบที่เกี่ยวข้องกับไฟล์ หน่วยความจำและสกรีน โครงสร้างการจัดการนี้หากพิจารณาเฉพาะอยู่ที่หน่วยความจำอย่างเดียว โดยถือว่าข้อมูลทั้งไฟล์ได้รับการโหลดมาที่หน่วยความจำแล้วก็จะทำให้การจัดการไม่ยาก แต่หากให้หน่วยความจำและไฟล์บนตัวกลาง เช่น ดิสค์มีความสัมพันธ์แบบ Virture disk จะทำให้โครงสร้างการจัดการยุ่งยากขึ้นอีกมาก

การจัดการข้อมูลเอกสารของเวิร์ดโปรเซสเซอร์ที่เป็นที่ต้องการพื้นฐานประกอบด้วย

- การเคลื่อนย้ายไปยังจุดที่ต้องการแก้ไข โดยสามารถเลื่อนตัวชี้ (file pointer) ไปยังตำแหน่งที่ต้องการได้รวดเร็ว
- ค้นหาข้อความตามสตริงที่กำหนด โดยค้นหาในรูปแบบต่าง ๆ เช่น ค้นหาทั้งหมดเอกสารหรือบางส่วนของเอกสาร
- บราวส์ เพื่อมองดูเอกสารทั้งหมดและสามารถแก้ไขได้
- ขอดูเอกสารเป็นบางส่วน โดยการเลือกเอกสารที่ต้องการ จัดการฟอร์แมตเอกสารเพื่อแสดงผลทางเอาต์พุต โดยกำหนดเอกสารให้ตรงกับฟิลิคัลเอาต์พุตได้



รูปที่ 4 โครงสร้างระบบเวิร์ดโปรเซสซิ่งในปัจจุบัน

- สร้างเอกสารให้ตรงตามเฟรมที่กำหนด เช่น ฟอร์ม จดหมาย สัญญา หรือแบบเสนอโครงร่าง โดยมีร่างของแบบฟอร์มปรากฏแล้วผู้ใช้เติมคำลงในช่องว่าง

- กำหนดขนาดหรือความกว้างของเอกสารได้อย่างอิสระ ทุกส่วนของเอกสาร

- เพิ่มลบข้อความของเอกสารบางส่วน โดยมีขีดความสามารถในการทำ Undo ได้ภายหลัง

- จัดการเอกสารเป็นบล็อก สามารถตัดต่อข้อความเสมือนใช้กรรไกรและเทปทำงานจริง

- มีขีดความสามารถในการจัดการชื่อเอกสารและรายละเอียดประกอบเพื่อให้ผู้ใช้ค้นหาเอกสารได้สะดวก และการจัดดัชนีชื่อเอกสารจะกระทำอย่างอัตโนมัติ

- มีคำสั่ง มาโคร หรือจัดทำแบบชวเลข หรือขีดความสามารถการทำงานแบบตารางได้

- มีการเสนอความช่วยเหลือแบบ online interactive help

ออฟไลน์:การจัดการเอกสารที่ยังมีใช้

อย่างไรก็ดี การประมวลผลหลายอย่างยังต้องใช้แบบออฟไลน์ เช่น การจัดการพิมพ์แบบ mailmerge หรือการจัดรูปแบบบางอย่างให้กับเครื่องพิมพ์ เช่น ในโปรแกรมบางโปรแกรมจะใช้ . เป็นตัวบอกคำสั่งออฟไลน์ เช่น

.ll - line length

.ds - double space

.pp - paragraph

เป็นต้น

การกำหนดรูปแบบคำสั่งนี้จะแตกต่างกันตามที่ผู้ออกแบบได้กำหนดขึ้น เช่น ในระบบ UNIX มีการกำหนดคำสั่งมาโคร เช่น

\ s8UNIX \ s10 runs on a \ s10 PDP \ s8 11/45

หมายความว่าให้พิมพ์ UNIX runs on a PDP 11/45

ตัวพิมพ์ ขนาด 8pt | ตัวพิมพ์ ขนาด 10pt | ตัวพิมพ์ ขนาด 10pt | ตัวพิมพ์ ขนาด 8pt

การนำเอาหลักการ NLP มาใช้ในเรื่องการลดขนาดเอกสาร

หลักการทางด้าน NLP มีบทบาทต่อการประมวลผลทางด้านคอมพิวเตอร์หลายอย่าง เช่น ในปัจจุบันเราจำเป็นต้องเก็บเอกสารข้อมูลเป็นจำนวนมาก และเพื่อให้การจัดเก็บมี

ประสิทธิภาพจำเป็นต้องอาศัยเทคนิคของการลดขนาดข้อมูลลง การลดขนาดนี้เป็นเรื่องที่อาศัยเทคนิคหลายรูปแบบ เช่น ค่าของสถิติการใช้ตัวอักษร การใช้คำ การใช้คู่ตัวอักษร เช่น จากการศึกษาของผู้เขียนพบว่าคำในภาษาไทยนั้น กว่า 80% ของเอกสารที่เราใช้จะใช้คำเพียงประมาณ 800 คำเท่านั้น หรือตัวอักษรภาษาอังกฤษนั้น 15 ตัวอักษรที่ใช้ คือ e,t,a,i,s,o,n,h,r และ d หากนับความถี่ของตัวอักษรเหล่านี้แล้วจะใช้ประมาณ 85 ถึง 95 เปอร์เซ็นต์ของเอกสาร สำหรับการลดขนาดเอกสารที่นิยมใช้ในขณะนี้วิธีหนึ่งคือการสร้างรหัส huffman โดยนำสถิติมาเป็นตัวกำหนดรหัส

สำหรับในทางปฏิบัติ รหัสแบบ huffman ไม่สะดวกต่อการประมวลผล เพราะความยาวไม่คงที่ทำให้การประมวลผลเป็นไปได้ยาก และการถอดรหัสก็ยุ่งยากด้วยอาจมีโอกาสผิดพลาดได้ง่าย จึงมีผู้คิดวิธีการสร้างรหัสแบบคงที่ โดยใช้คู่ตัวอักษรที่ใช้มากมาอยู่ในตารางด้วยเช่น ตัวรหัสภาษาอังกฤษที่ใช้ในเอกสารประกอบด้วย

ตัวช่องว่าง (blank)	1
ตัวอักษรเล็ก	26
ตัวอักษรใหญ่	26
ตัวเลข	10
เครื่องหมายพิเศษ	25
รวม	88

แต่รหัสที่ใช้ขนาด 8 บิต แทนได้ 256 รหัส ดังนั้นส่วนที่เหลือจึงเป็นรหัสคู่ หรือรวมตัวอักษรหลาย ๆ ตัว เช่น เมื่อเก็บคำว่า among the people จะเก็บในรูปแบบดังรูปที่ 5

AM	ON	G	#T	H	E#	P	EO	P	L	E#

รูปที่ 5 เครื่องหมาย # คือช่องว่าง การเก็บรหัสคู่แบบนี้ทำให้การประมวลผลทำได้ง่ายมาก

เครื่องหมาย # คือ ช่องว่าง การเก็บรหัสคู่แบบนี้ทำให้การประมวลผลทำได้ง่ายมาก

การเอ็นคริพชัน

การประมวลผลข้อความอีกประการหนึ่งคือ การแปลงรหัสตัวอักษรให้อยู่ในรูปแบบที่ผู้อื่นจะไม่ทราบได้ว่าประกอบด้วยข้อมูลอะไร ทั้งนี้เพราะการที่จำเป็นต้องมีการส่งถ่ายข้อมูลระหว่างกัน ข้อมูลบางอย่างอาจต้องกำหนดให้เป็นความลับ ดังนั้น จึงต้องมีการกำหนดรหัสในรูปแบบเอ็นคริพชันเพื่อแปลงให้อยู่ในรูปที่ยากต่อการคาดเดาหรือถอดได้โดยง่าย เทคโนโลยีของการเอ็นคริพชัน และดีคริพชันจึงเป็นศาสตร์ที่มีการพัฒนาและใช้หลักการของ NLP ประกอบด้วย

ภาษาศาสตร์คอมพิวเตอร์กับ NLP

การประมวลผลภาษาธรรมชาติเป็นความพยายามของมนุษย์ที่จะหาข้อมูลและข่าวสารที่เกี่ยวกับภาษาธรรมชาติ โดยเห็นปัญหาของการใช้และการแทนภาษาธรรมชาติในรูปแบบที่คอมพิวเตอร์จะประมวลผลได้ ดังนั้นศาสตร์ทางด้าน NLP จึงเป็นศาสตร์ที่จะให้คอมพิวเตอร์เข้าใจภาษามนุษย์ NLP จึงเป็นศาสตร์ที่เกี่ยวกับการแทนความรอบรู้ โดยอาศัยหลักการทางด้านปัญญาประดิษฐ์ โดยใช้ความรอบรู้ทางภาษาศาสตร์เป็นหลัก อย่างไรก็ตาม NLP ยังเกี่ยวข้องกับเทคนิคทางคอมพิวเตอร์และวิทยาศาสตร์ในเรื่องของ cognitive science

ในการประมวลผลภาษาธรรมชาตินั้นเกี่ยวข้องกับหลักการ ดังนี้

การสังเคราะห์ ซึ่งได้แก่ การสร้างโครงสร้างของภาษาจากหลักการและโครงสร้างที่เก็บไว้ทั้งรูปแบบความหมาย ไวยากรณ์และคำ

การวิเคราะห์ ซึ่งได้แก่ การแยกแยะเนื้อหากฎเกณฑ์บางอย่าง เพื่อนำไปใช้หรือกำหนดเป็นหลักการ สิ่งสำคัญคือจะหาวิธีการแยกแยะให้เป็นหน่วยย่อยที่มีความสอดคล้องกัน

การ transformation คือ การเปลี่ยนรูปจากแบบหนึ่งไปเป็นอีกแบบหนึ่งที่สมนัยกัน

การ inference คือ การถ่ายทอดความรู้จากระบบเพื่อนำไปใช้ในการสร้างหรือประยุกต์

การประมวลผลภาษาธรรมชาติจึงเป็นความสามารถที่มนุษย์จะให้เครื่องจักรเข้าใจมนุษย์ได้ดีขึ้น นับว่าเป็นการท้าทายยิ่งนัก

การศึกษาภาษาศาสตร์ เริ่มจากต้องแบ่งชนิดของคำ

Zellig S. Harris นักภาษาศาสตร์ได้เสนอกฎเกณฑ์การแทนที่ไว้ในปี 1943 โดยแฮริสเสนอว่าการศึกษาภาษาศาสตร์จำเป็นต้องมีการแบ่งชนิดของคำ เพื่อที่จะนำมาทดแทนกันในกลุ่มคำชนิดเดียวกัน เช่น ประโยคที่เรียงคำเป็น C A D และ C B D ถ้า A และ B อยู่ในกลุ่มคำชนิดเดียวกับ A และ B จะใช้แทนกันในการสร้างประโยคได้ ลองพิจารณาจากตัวอย่างประโยค

Where did the....go?

สิ่งที่เว้นไว้นี้สามารถหาคำอะไรก็ได้ที่อยู่ในชนิดคำแบบเดียวกัน เช่น จะแทนที่ child หรือ young boy

ดังนั้น การศึกษาภาษาศาสตร์เพื่อให้ประมวลผลได้จำเป็นจะต้องจัด class ของคำ เช่น ประโยค The monkey has

eaten the banana เราก็จะต้องแบ่งว่าแต่ละคำอยู่ใน class ใด เช่น เป็น นาม กริยา เป็นต้น

อย่างไรก็ตามเชริสยังเสนอว่าประโยคที่ใช้กันนี้ยังสามารถแปลงรูปโครงสร้างไปมาได้ (transform) การแปลงรูปนี้จะเปลี่ยนไปทั้งโครงสร้าง เช่น การเปลี่ยนระหว่าง active voice กับ passive voice หรือเปลี่ยนรูปจากบอกเล่าเป็นประโยคคำถาม เช่น

The dog bit the man.

The man was bitten by the dog.

หรือ

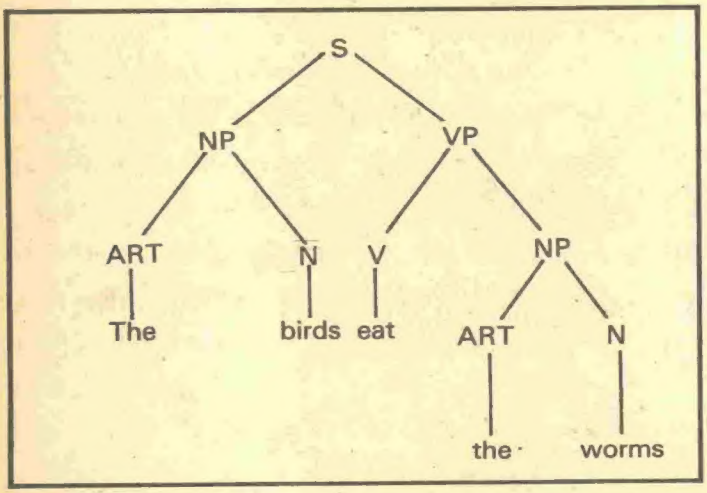
The man was bitten by the dog.

Was the man bitten by the dog?

NOAM CHOMSKY ผู้ชี้แนวทางการประมวลผลภาษาธรรมชาติ

ในปี 1957 ชมสกี้ได้ตีพิมพ์หนังสือขึ้นมาเล่มหนึ่งให้ชื่อว่า "Syntactic Structure" หนังสือเล่มนี้ถือว่าเป็นจุดเริ่มต้นของการชี้แนวทางของการศึกษาภาษาศาสตร์ในแนวใหม่ที่จะนำมาใช้ในการประมวลผลได้ ชมสกี้ได้เสนอวิธีการแทนรูปไวยากรณ์ด้วยการสร้างเป็นกฎเกณฑ์ที่แน่นอนขึ้น เพื่อให้การวิเคราะห์หรือการสร้างประโยคเป็นไปอย่างชัดเจนและต่อมาได้ตั้งชื่อว่า PS grammar (phase structure grammar) ตัวอย่างการกำหนดไวยากรณ์เป็นแบบการกำหนดเฉพาะลงไปด้วยกฎเกณฑ์ เช่น

S → NP + VP
NP → ART + N
VP → V + NP



รูปที่ 6 โครงสร้างของการแตกกระจายประโยค the birds eat the worms

N → birds, worms, boys, cars, rocks, ...
ART → the
V → eat, drive, learn

ซึ่งเมื่อเขียนเป็นรูปโครงสร้างของไวยากรณ์ตามที่กำหนดเพื่อแยกประโยค The birds eat the worms ได้ดังรูปที่ 6

อย่างไรก็ดีตัวอย่างที่ออกมานี้เป็นตัวอย่างแบบง่าย ๆ ไวยากรณ์ PS นี้ยังมีโครงสร้างที่เป็นแบบรีเคอร์ซีฟได้ ดังนั้นโครงสร้างไวยากรณ์ที่จะเขียนสำหรับประโยคที่ซับซ้อนก็จะทำได้เช่นเดียวกัน

ในเวลาต่อมา ชมสกี้ได้สร้างทฤษฎีมาตรฐานที่เรียกว่า standard theory ซึ่งเป็นทฤษฎีแม่บทของความสัมพันธ์ของโครงสร้างรูปลึก (deep structure) กับโครงสร้างรูปผิว (surface structure) ตลอดจนการสร้างเป็นคำพูดและความหมาย ทฤษฎีที่เน้นให้เห็นโครงร่างของกรรมวิธี NLP ที่จะต้องเกี่ยวข้องกับข่ายต่าง ๆ ของระบบความสัมพันธ์เพื่อจะให้เครื่องจักรมีความคิดอย่างมนุษย์ ทฤษฎีนี้ได้รับการขยายต่อกลายเป็น extended standard theory ซึ่งเขียนเป็นรูปไดอะแกรมได้ดังรูปที่ 7

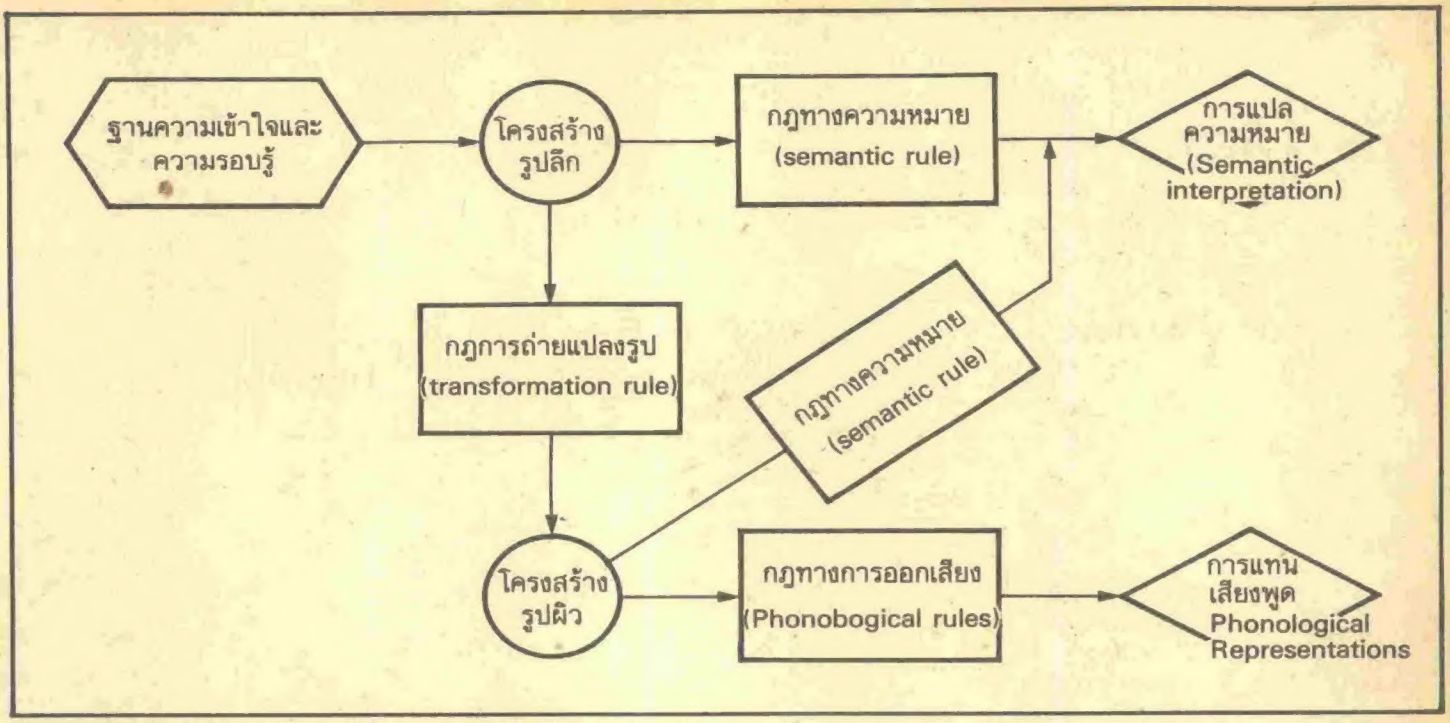
ด้วยหลักการของไวยากรณ์ PS ทำให้ดูการใช้ภาษาเป็นโครงสร้างขึ้นพอที่จะประมวลผลได้ แต่อย่างไรก็ตามปรากฏว่าวิธีการวิเคราะห์โครงสร้างไวยากรณ์ PS นี้ยังให้ความกำกวมอยู่มากและความยุ่งยากของการวิเคราะห์จะเพิ่มมากขึ้นถ้าโครงสร้างประโยคนั้นยุ่งยากขึ้น ด้วยเหตุนี้เองจึงเป็นอุปสรรคของงานประมวลผล NLP ในยุคต่อมา จึงต้องมีผู้คิดพยายามหารูปแบบไวยากรณ์รูปอื่นมาใช้

ไวยากรณ์การก (Case grammar) ความหวังของการสร้างโครงร่างสากล

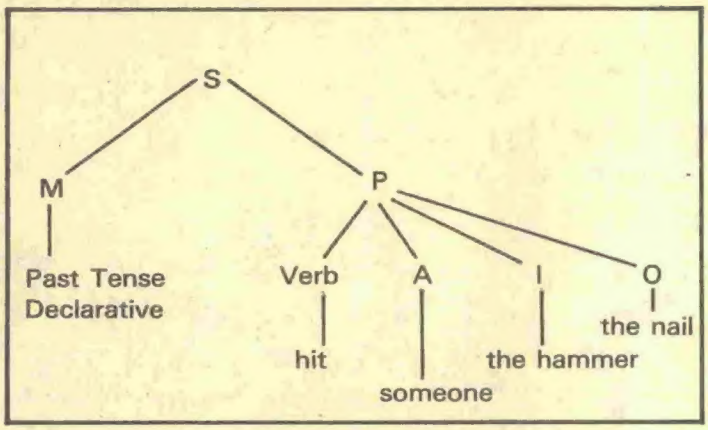
ในปี ค.ศ. 1967 ชาร์ล ฟิลมอร์ ได้เสนอหลักการโครงสร้างไวยากรณ์แบบใหม่ โดยแยกโครงร่างของประโยคออกเป็นสองส่วน ส่วนหนึ่งเป็นส่วนที่ค่อนข้างจะอยู่ตัวหรือเรียกว่าสแตติก กับอีกส่วนเป็นส่วนประกอบที่ทำให้ประโยคมีความอ่อนตัวมาก ไวยากรณ์การกแบ่งโครงร่างประโยคเป็นสองส่วนเรียกว่า Modality และ Proposition

ส่วน Modality เปรียบเสมือนเป็นแอตทริบิวต์ของประโยค โดยส่วนนี้จะแสดงปรากฏการณ์ทางรูปผิวของประโยคหลายอย่าง เช่น เรื่องของ tense, aspect, form, mood, essence, modal, manner และ time

ส่วน Preposition เป็นส่วนที่แสดงความสัมพันธ์ของกริยากับนาม หรือส่วนแสดงความสัมพันธ์ของแอคชันกับผู้ที่หรือสิ่งที่เกี่ยวข้องด้วย



รูปที่ 7 ทฤษฎีมาตรฐานของชมสกี้ที่ขยายเพิ่มขึ้น



รูปที่ 8 การกระจายตามไวยากรณ์การก

โครงสร้างของไวยากรณ์การก เขียนได้เป็น

$$S \longrightarrow M + P$$

$$M \longrightarrow \text{Tense, aspect....}$$

$$P \longrightarrow C1 + C2 + C3 \ldots Cn$$

โดยในสมัยแรก ฟิลมอร์ได้กำหนดไว้เพียง 6 การก คือ Agentive, Instrument, Dative, Factitive, Locative และ Objective

เช่นประโยค someone hit the nail with the hammer เมื่อกระจายตามโครงร่างไวยากรณ์การก จะได้เป็นดังรูปที่ 8

สังเกตว่าไวยากรณ์การกนี้ได้แยกโครงร่างประโยคที่แสดงความสัมพันธ์ของกริยากับคำนามหรือนามวลี ด้วยเหตุนี้เองจึงเกี่ยวข้องลงไปในเชิงความหมายอยู่ด้วย และการแยกโครงร่างไวยากรณ์การกในลักษณะนี้ทำให้เป็นแบบสากลที่ทุกภาษาสามารถใช้ได้เหมือนกัน ดังนั้นโครงการแปลภาษาด้วยเครื่องของหลายบริษัทในประเทศญี่ปุ่นได้ใช้หลักการไวยากรณ์การกนี้เป็นหลัก และจัดโครงสร้างของข้อมูลแทนภาษากลาง-ว่า interlingua

อย่างไรก็ตาม การกำหนดจำนวนการกนี้ในแต่ละระบบได้เพิ่มเติมเปลี่ยนแปลงเข้าไปจนทำให้โครงร่างของการกที่ใช้มีมากกว่า 20 การก ซึ่งก็เป็นปัญหายุ่งยากต่อระบบ parser เช่นกันที่จะทำการแปลงรูปประโยคให้เข้าสู่รูปโครงของไวยา-กรณ์ดังกล่าว

การแทนรูปโครงไวยากรณ์ด้วยกราฟ

การประมวลผลภาษาธรรมชาตินี้เป็นการหาหลักการที่จะทำการประมวลผลด้วยคอมพิวเตอร์ได้ ดังนั้น จึงมีผู้พยายามคิดค้นวิธีการที่จะใช้การแทนไวยากรณ์แบบต่าง ๆ วิธีหนึ่งที่ค่อนข้างประสบผลสำเร็จ คือ การกำหนดในรูปแบบกราฟที่เรียกว่า finite state transition graph เช่น กลุ่มประโยคต่อไปนี้ คือ

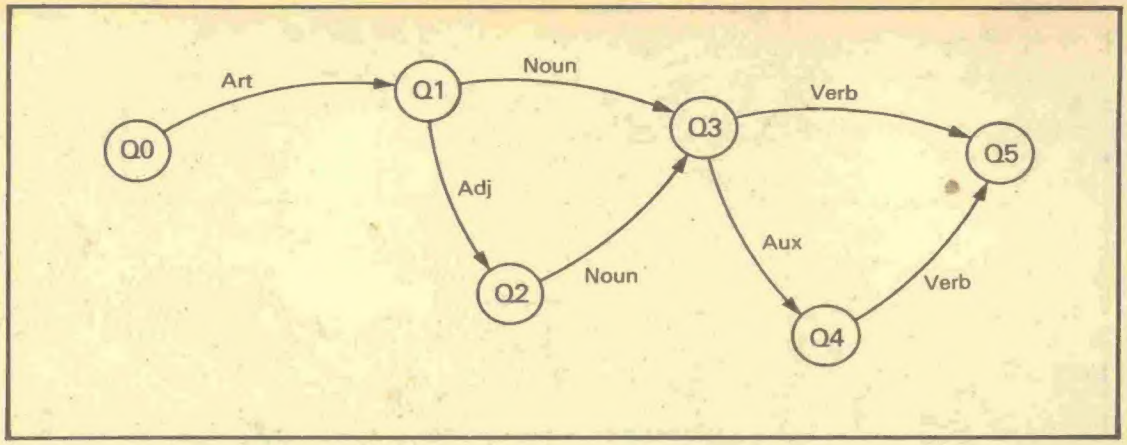
The boy may swim.

The brown dog run.

A bad girl cries.

The little children can sing.

รูปที่ 9
การแทนไวยากรณ์
ด้วยกราฟ



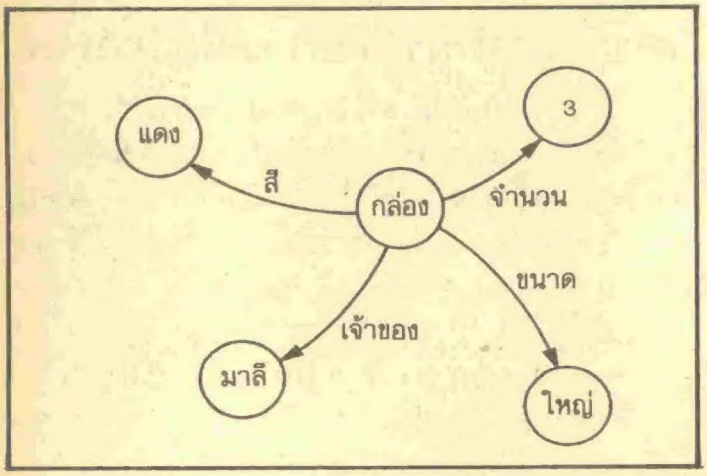
พิจารณาโครงสร้างประโยคแรกประกอบด้วย Article, Noun, Auxilliary, Verb

เมื่อเขียนในรูปกราฟที่จะแทนประโยคทั้ง 4 จะเขียนได้เป็นดังรูปที่ 9

ส่วนของเน็ตเวอร์กที่ใช้แทนนี้ยังสามารถรีเคอร์ซีฟได้ และเมื่อเป็นเช่นนี้จึงมีผู้พัฒนาเป็นเน็ตเวอร์กแบบ RTN - Recursive Transitions Network และต่อมาพัฒนาขึ้นเป็น ATN-Augmentation Transition Network ซึ่งรายละเอียดดังกล่าวนี้ ผู้สนใจหาอ่านได้จากบทความที่ชื่อ Transition Network Grammar for Natural lanquage Analysis โดย William A.Woods ในหนังสือ Comme ACM Vol 13 October 1970

ยุคของ AI เป็นยุคของการแทนความรอบรู้

การประมวลผลภาษาธรรมชาติมาตื่นตัวมากอีกครั้งหนึ่งในยุคของเอไอที่กำลังเบ่งบาน ปรากฏว่ามีผู้สร้างระบบที่จะใช้แทนความรอบรู้ในระดับความหมายของคำอยู่มาก การแทนความหมายนี้เป็นวิธีหนึ่งของการแทนความรอบรู้ เช่น จากประโยค "กล่องสีแดงขนาดใหญ่สามกล่องของมาลี" สามารถเขียนแทนเป็นไดอะแกรมแสดงความสัมพันธ์ได้ดังรูปที่ 10



รูปที่ 10 แสดงความสัมพันธ์ในเชิงความหมาย

และด้วยหลักการของเอไอนี้ก็มีการพัฒนาระบบคณิตศาสตร์เชิงลอจิกที่เรียกว่า predicate calculus ขึ้นมา และทำให้คณิตศาสตร์แขนงนี้ใช้ในการแทนความรอบรู้และแก้ปัญหาเชิงลอจิกได้มาก การแทนความหมายเชิงลอจิกนี้สามารถสร้างเป็นเน็ตเวอร์กขึ้นมาที่เราเรียกว่า ซีแมนติกเน็ตเวอร์ก เช่น จากรูปที่ 10 เป็นการแสดงโครงสร้างความสัมพันธ์ของโมเดลความสัมพันธ์ที่สร้างเป็นเน็ตเวอร์กขึ้นมา และสามารถแทนในรูปเชิงคณิตศาสตร์ลอจิก หรือ predicate calculus ได้รูปที่ 11 เป็นตัวอย่างของซีแมนติกเน็ตเวอร์ก

จากเน็ตเวอร์กดังกล่าวบอกได้ว่า ปีติอยู่ในโลก สิ่งที่เป็นสัตว์ (animal) ได้แก่ piti, noy, kim เป็นต้น จะเห็นว่าจากโครงสร้างดังกล่าวเราสามารถนำมาใช้ในการแทนความหมายของสิ่งต่าง ๆ ได้ และสามารถใช้หลักการทางคณิตศาสตร์ลอจิกเพื่อค้นหาความจริงได้

ทฤษฎี Conceptual Dependency เป็นทฤษฎีที่ใช้อ้างอิงหลักการแทนความหมาย

โรเจอร์ แชงค์ (Roger Schank) เป็นนักภาษาศาสตร์คนหนึ่งที่ได้รับการกล่าวถึงและอ้างอิงกันมากในยุคของการแทนความรอบรู้ที่เกี่ยวกับความหมายอยู่มาก กล่าวคือคำทุกคำน่าจะมี concept ของความหมายอยู่ เราจะหาวิธีแทน concept เหล่านี้ได้อย่างไร แชงค์ได้เสนอหลักการของการสร้าง primitive act ซึ่งว่าด้วยการแทน concept ในกริยาต่าง ๆ เพื่อนำมาประกอบกัน การแสดงความหมายต่าง ๆ โดย primitive act ของแชงค์ประกอบด้วย

MOVE คือการเคลื่อนไหวของส่วนหนึ่งส่วนใดของสิ่งมีชีวิต

PROPEL คือการกระทำโดยบุคคลที่ทำให้วัตถุเปลี่ยนตำแหน่ง

INGEST คือการเกี่ยวข้องการกระทำในการนำบางสิ่งบางอย่างเข้าไปในร่างกายของสิ่งมีชีวิตเพื่อใช้ในการอธิบายการกิน การดื่ม

GRASP คือการกระทำของบุคคลที่ทำการหยิบจับสิ่งของ

EXPEL คือการกระทำที่ตรงกันข้ามกับ INGEST

ATRANS คือการเคลื่อนย้ายวัตถุ ในลักษณะ abstract

PTRANS คือการเคลื่อนย้ายวัตถุแบบ physical

ATTEND ใช้สำหรับในกรณีที่บุคคลรับข้อมูลข่าวสารจากแหล่งภายนอก

MTRANS คือการเคลื่อนย้ายข้อมูลข่าวสาร ซึ่งเป็นลักษณะ mental transfer

MBUILD เกี่ยวข้องกับการเชื่อมโยงของ mental information ที่ใช้ใน concept ที่เกี่ยวกับการคิดหรือตัดสินใจ

การแทน primitive act นี้นำมาใช้แทนความหมายทาง concept ต่าง ๆ เช่น ประโยค "มาลีให้หนังสือปิติ" เขียนในรูปแบบตามทฤษฎี conceptual dependency ได้เป็น

EVENT1
ACTOR: มาลี
ACTION: ATRANS
OBJECT: หนังสือ
DIRECTION: FROM: มาลี
TO: ปิติ

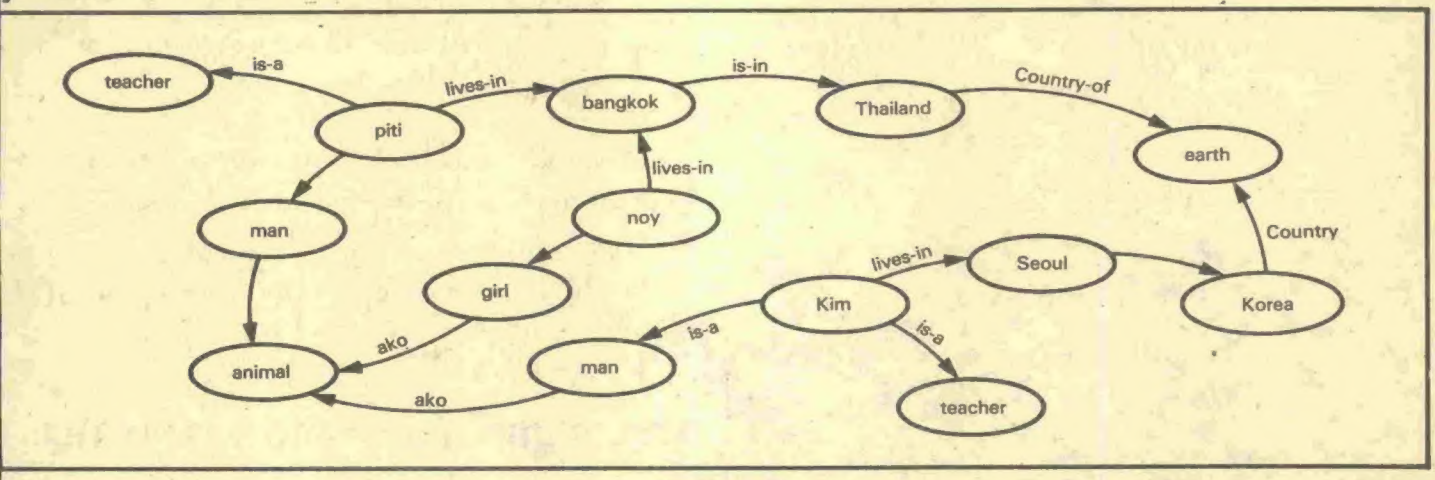
หรือจากประโยค "มานะอยู่ที่เชียงใหม่" เขียนเป็น

SITUATION1
OBJECT: มานะ
STATE: ตำแหน่ง
VALUE: เชียงใหม่

ด้วยการสร้างทฤษฎี CD (Conceptual Dependency) แชงค์ได้เขียนหนังสือเรื่อง script and plan ซึ่งทำด้วยการใช้ primitive act บรรยายความหมายของข้อความ หรือเรื่องราวต่าง ๆ เพื่อสร้างโมเดลที่แทนการกระทำหรือสามารถแทนความรอบรู้ เพื่อใช้ในการประมวลผลด้วยคอมพิวเตอร์ได้

ตัวอย่างสคริปต์ที่ใช้แทนเช่น สันติเข้าไปในภัตตาคารและสั่งอาหารทาน ซึ่งสามารถเขียนเป็นฉากของสคริปต์ที่แทนโครงเรื่องทั้งหมดได้ดังรูปที่ 12

รูปที่ 11 แสดงซีแมนติกเน็ตเวอร์ก



รูปที่ 12 ตัวอย่างสคริปต์ที่ใช้แทนความรอบรู้

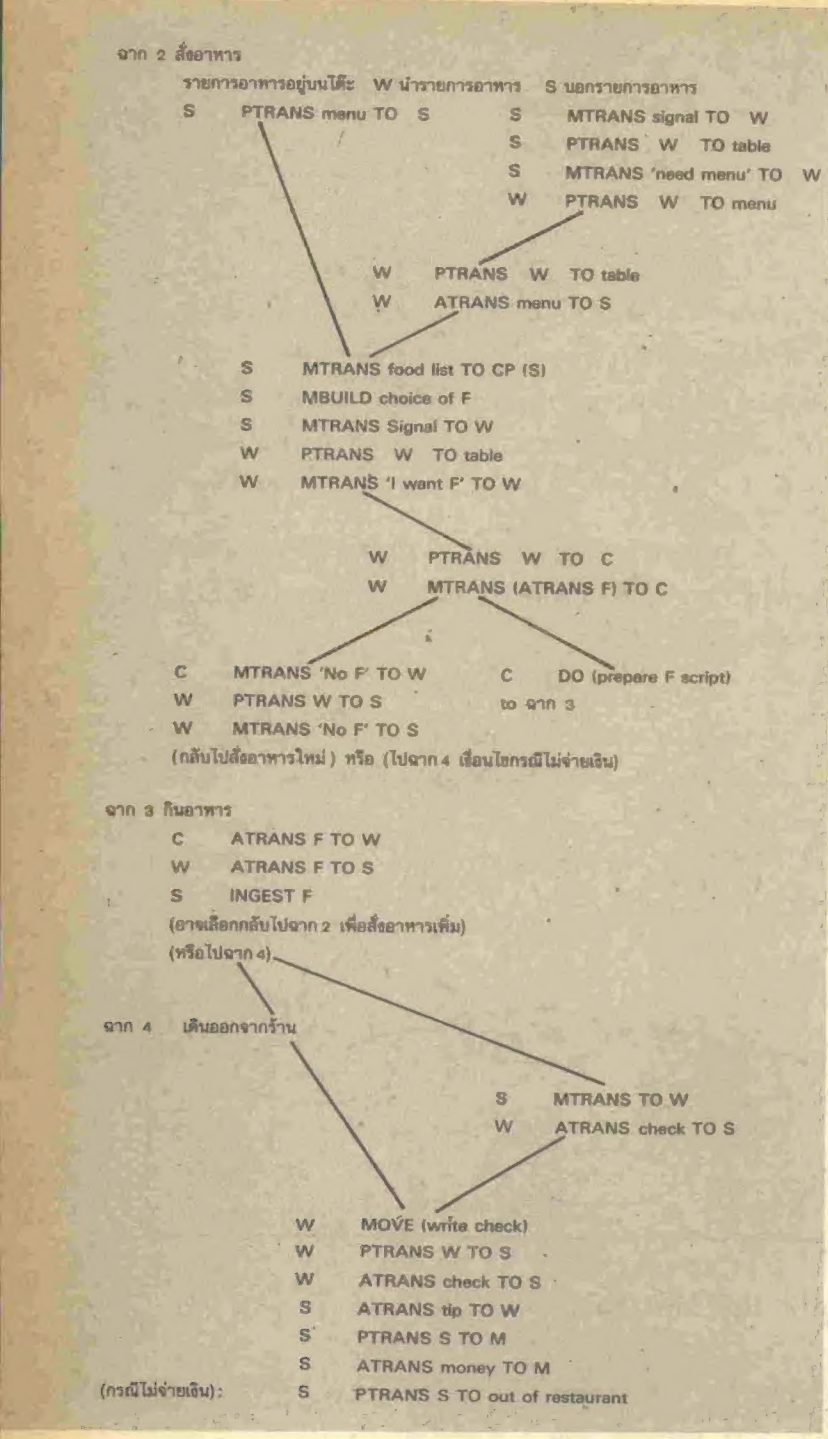
สคริปต์: ภัตตาคาร
ฉาก: โต๊ะ (Tables)
รายการอาหาร (Menu)
อาหาร (F-Food)
เงิน (Money)

ตัวแสดง: S-ลูกค้า
W-พนักงานเสิร์ฟ
C-คนครัว
M-คนเก็บเงิน
O-เจ้าของร้าน

เงื่อนไขเริ่มต้น S หิว
S มีเงิน

ผลลัพธ์ S มีเงินลดลง
O มีเงินเพิ่มขึ้น
S อิ่ม (หายหิว)
S พอใจ

ฉาก 1 การเดินเข้าร้าน
S PTRANS S into restaurant
S ATTEND eyes TO tables
S MBUILD where to sit
S PTRANS S TO table
S MOVE S TO sitting position



สังเกตจากโครงเรื่องที่เขียนดังกล่าวนี้แทนความรอบรู้ที่เราอาศัย primitive act ที่แชงค์เป็นผู้กำหนดขึ้น ด้วยวิธีนี้เราสามารถแทนความรอบรู้เพื่อการประมวลผลในระดับความหมายได้

พจนานุกรมอิเล็กทรอนิกส์ จึงเป็นข้อมูลสำคัญของงาน NLP

ความจริงเกี่ยวกับภาษาเป็นเรื่องที่ต้องเก็บรวบรวมข้อมูลไว้เพื่อให้คอมพิวเตอร์เรียกใช้ พจนานุกรมอิเล็กทรอนิกส์จึงเสมือนแหล่งรวมข้อมูลระดับมหาศาลไว้สำหรับการประมวลผล ข้อมูลที่ทำการจัดเก็บในพจนานุกรมอิเล็กทรอนิกส์จะประกอบด้วย

- คำหรือหน่วยคำ ซึ่งจัดเป็น entry word
- ข้อมูลระดับคำ และการแบ่งแยกคำและพยางค์
- ข้อมูลเกี่ยวกับตัวสะกด ตัวย่อ
- รายละเอียดเกี่ยวกับการผันแปรของคำ
- การออกเสียง
- ชนิดของคำ (part of speech)
- ความถี่ที่ใช้ และอัตราความสำคัญของคำ
- ข้อมูลเกี่ยวกับลักษณะของคำที่เป็นคำเก่าแก่ คำใหม่ คำที่ใช้ทั่วไป ราชาศัพท์ ฯลฯ
- โครงสร้างทางไวยากรณ์ที่เกี่ยวข้อง
- เฟรมของการก (Case frame)
- ตัวอย่างการใช้และประโยค ตลอดจนรูปแบบประโยคที่ใช้
- คำพ้องความหมาย (thesaurus)
- คำแปลในต่างภาษา
- การเชื่อมโยงในระดับ Concept และไวยากรณ์
- ฯลฯ

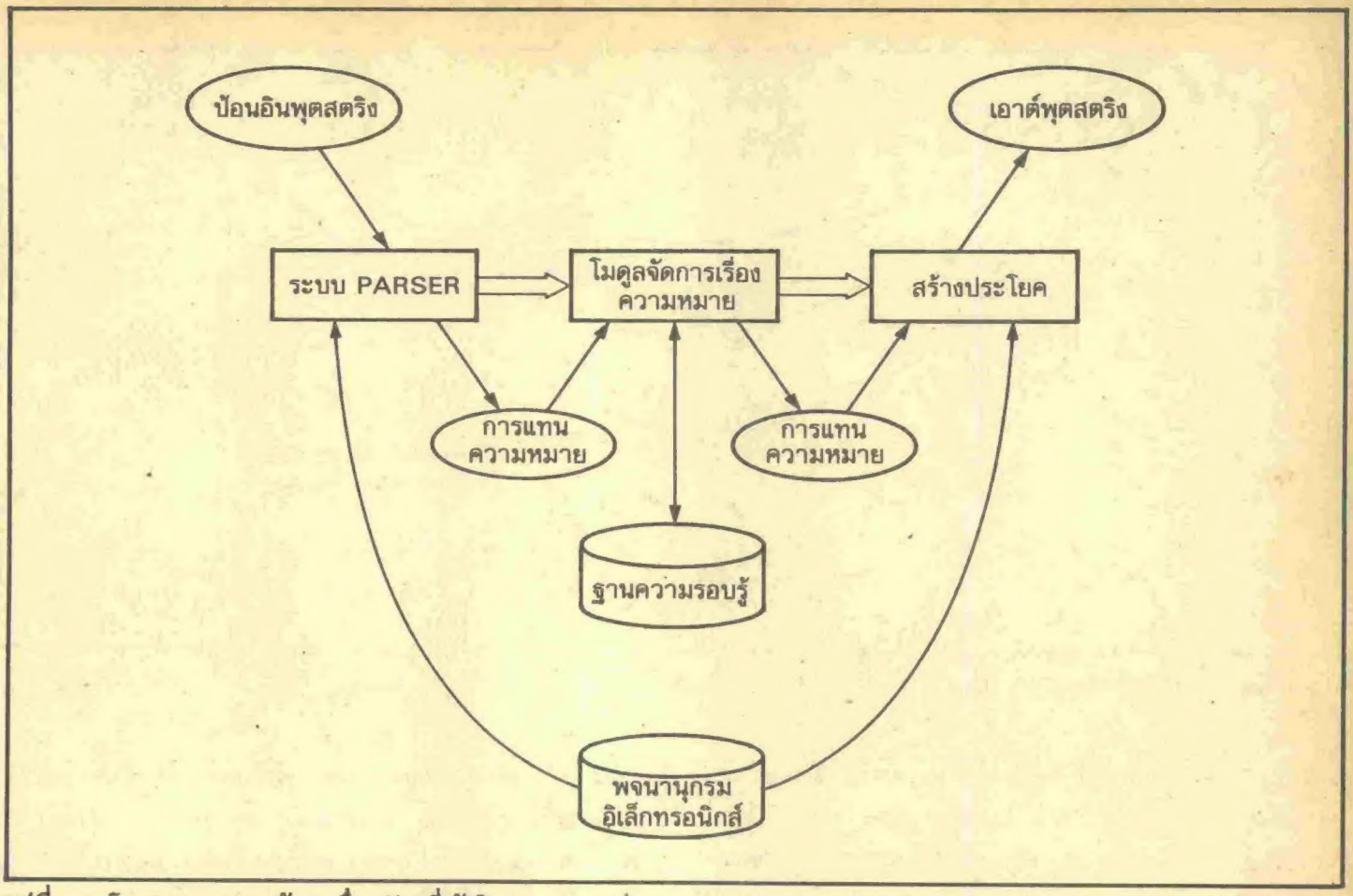
ผู้เขียนมีความสนใจในเรื่องการเก็บรวบรวมข้อมูลทางด้านภาษาศาสตร์ของภาษาไทย และมีความเห็นว่าหากเรามีการรวบรวมข้อมูลทางด้านภาษาไทยไว้เป็นฐานข้อมูล ก็จะทำให้งานประมวลผลภาษาธรรมชาติที่จะมีขึ้นในอนาคตเป็นไปได้ดียิ่งขึ้น

การประยุกต์ให้เครื่องจักรเข้าใจภาษาธรรมชาติ

ผู้ที่ศึกษาทางด้านเอไอจะพบว่ามีตัวอย่างโปรแกรมที่มีชื่อเสียงหลายโปรแกรมเช่น ELIZA เป็นตัวอย่างของการโต้ตอบระบบทาง man-machine ในลักษณะภาษาธรรมชาติ อย่างไรก็ตาม การตอบนั้น ELIZA ก็จะตอบในขอบเขตของฐานความรอบรู้ที่จำกัดมาก

โมเดลของการทำให้เครื่องจักรเข้าใจภาษาธรรมชาติได้เขียนได้ดังรูปที่ 13

เมื่อป้อนข้อความที่เป็นภาษาธรรมชาติมา ประโยคหรือข้อความนั้นจะได้รับการนำมาวิเคราะห์ทางไวยากรณ์ก่อน นั่นคือ ผ่านขบวนการ parser ประโยค การทำการวิเคราะห์ไวยากรณ์นี้จำเป็นต้องอาศัยข้อมูลจากพจนานุกรมอิเล็กทรอนิกส์ เมื่อแยกแยะไวยากรณ์ได้แล้วก็จำเป็นต้องหาความหมายของข้อความขึ้น โดยจัดเข้าสู่โมดูลที่จะทำการค้นหารายละเอียด



รูปที่ 13 โมเดลของการสร้างเครื่องจักรที่เข้าใจภาษามนุษย์

ที่เกี่ยวข้องในฐานความรอบรู้ เมื่อหาข้อมูลจากฐานความรอบรู้ที่เกี่ยวข้องได้แล้วก็จะจัดเตรียมโครงสร้างที่จะสร้างประโยคออกเป็นสตริงที่เอาต์พุต

การแปลภาษาด้วยคอมพิวเตอร์ : โครงงานที่ท้าทายความสามารถมนุษย์

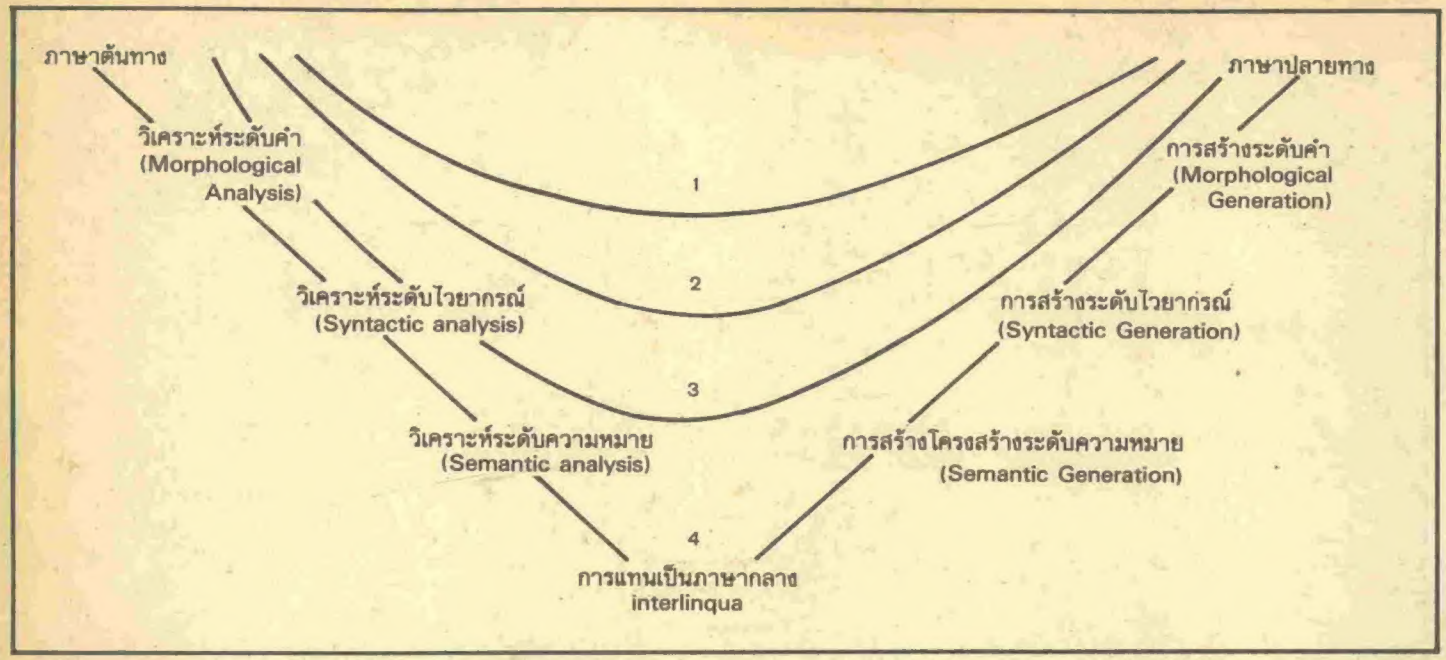
จากโครงงานการแปลภาษาที่ผู้เขียนได้มีโอกาสร่วมอยู่ด้วย ตั้งแต่ครั้งโครงงานที่ทดลองทำด้วยโมเดลเล็ก ๆ ในห้องปฏิบัติการ และจากการศึกษาโมเดลการแปลของอีกหลาย ๆ รูปแบบ พบว่า การแปลภาษาด้วยคอมพิวเตอร์นั้นเป็นเรื่องที่ต้องอาศัยศาสตร์หลายแขนงเพื่อประกอบกันเป็นงานสร้างชิ้นที่สำคัญ การแปลภาษาจะต้องใช้เทคโนโลยีทางด้านการจัดการข้อมูลที่ดีเยี่ยม ใช้หลักการแทนความรอบรู้ที่สามารถสร้างงานความรอบรู้ขนาดใหญ่ และนำฐานความรอบรู้นั้นมาใช้ในการตัดสินแก้ไขปัญหาด้วยคอมพิวเตอร์ด้วยดี ประกอบกับต้องมีความรู้ทางด้านภาษาศาสตร์ที่ดีด้วย

โมเดลของการแปลภาษาด้วยเครื่องนั้นอาจแยกแยะตามระดับการถ่ายความหมายจากภาษาหนึ่งไปยังอีกภาษาหนึ่ง ซึ่งเมื่อเขียนเป็นรูปงานแสดงได้ดังรูปที่ 14

หากพิจารณาจากรูปโมเดลการแปลจะเห็นว่าระบบการแปลเราแบ่งแยกออกเป็น 4 วิธี ที่จะใช้ในการแปลได้

วิธีแรก เป็นวิธีที่หากมีคู่ศัพท์จำนวนน้อย และได้รับการกำหนดเป็นคีย์เวิร์ดที่แน่นอน มีขอบเขตการแปลที่จำกัดมาก เช่น ในคู่มือนักท่องเที่ยวทำเป็นระบบแปลแบบเครื่องคิดเลขเล็ก ๆ และเมื่อคีย์คำว่า How many? ก็จะแปลว่าเท่าไร เป็นต้น คีย์เวิร์ดของการแปลจึงอยู่ในขอบเขตของการใช้งานจำกัด ระบบการแปลในลักษณะนี้มีผู้ทำเป็นสินค้าขายแล้ว เช่น การแปลภาษาอังกฤษเป็นญี่ปุ่นสำหรับนักท่องเที่ยว เป็นต้น จะเห็นว่าการถ่ายจากภาษาต้นฉบับไปเป็นภาษาปลายทางทำได้ในระดับคำ

วิธีที่ 2 เป็นการ transfer ในระดับไวยากรณ์ วิธีนี้มีผู้พยายามทำขึ้นและเขียนเป็นหนังสือออกมาและแนะนำว่าเป็นวิธีการแปลที่ง่ายอีกวิธีหนึ่ง เช่น ในประเทศอังกฤษมีการตั้งกลุ่มที่ชื่อว่า SLUNT และเขียนโปรแกรมเป็นตัวอย่างการแปลภาษาอังกฤษกับฝรั่งเศส การวิเคราะห์จะใช้โครงสร้างตัวเลขแทนโครงสร้างไวยากรณ์ภาษาอังกฤษ เมื่อวิเคราะห์ไวยากรณ์และแทนตัวเลขได้ครบหมดแล้วก็จะนำข้อมูลตัวเลขไปหาในดิกชันนารี เพื่อนำมาแทนแล้วแปลงโครงสร้างตัวเลข



รูปที่ 14 โมเดลของระบบการแปลภาษาด้วยเครื่องคอมพิวเตอร์

กลับเป็นภาษาปลายทาง ผู้ที่พยายามแนะนำวิธีนี้และเสนอตัวอย่าง คือ W.Goshawke และคณะ ผู้สนใจอาจหาได้จากหนังสือชื่อ "Computer Translation" วิธีการแปลในลักษณะนี้ยังมีจุดอ่อนอีกมากมาย (ผู้เขียน)

วิธีที่ 3 เป็นระบบที่เรียกว่า transfer approach เป็นระบบที่มีการทำกันมากและทำกันมานานแล้ว คือ จะวิเคราะห์เข้าไปในระดับความหมาย แต่โครงสร้างของความหมายยังไม่มีการแตกกระจายโครงรูปร่างมากนัก โดยยึดหลักความหมายที่เป็นไปได้ในพจนานุกรม และเลือกคู่ความหมายที่สัมพันธ์กันในประโยคที่คิดว่าดีที่สุดในการจ่ายเข้าสู่ภาษาปลายทาง

วิธีที่ 4 เรียกว่า อินเตอริงกัว หรือเป็นการสร้างภาษากลาง ภาษากลางคืออะไร ภาษากลางคือ ข่ายของการแทน concept ของความหมายต่าง ๆ ที่มีความสัมพันธ์กันเป็นเน็ตเวอร์กเพื่อที่จะเหลือเฉพาะรูปแทนความหมายอย่างเดียว โดยโครงสร้างของส่วนประกอบประโยคจะถูกเก็บมาเป็นแอตทริบิวต์ของภาษากลาง เช่น modality การจัดโครงสร้างภาษากลางจึงเป็นโมเดลเชื่อมกันด้วยความสัมพันธ์ และความสัมพันธ์ส่วนหนึ่งจะใช้ไวยากรณ์การก อินเตอริงกัวนี้เป็นระบบแปลที่รัฐบาลญี่ปุ่นร่วมกับอีก 4 ชาติในเอเชีย ซึ่งรวมไทยดำเนินโครงการอยู่ ผลการดำเนินนี้ใช้เวลาทั้งสิ้น 6 ปี และโครงการได้ผ่านมาแล้ว 2 ปี ปัจจุบันโครงการกำลังอยู่ในระหว่างทดลองโมเดลขนาดเล็ก โดยมี 4 มหาวิทยาลัยไทยเกี่ยวข้อง คือ มหาวิทยาลัยเกษตรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าฯ ลาดกระบังและธนบุรี ร่วมอยู่ในโครงการ

อนึ่ง สำหรับเรื่องราวของ MT นั้น ผู้เขียนได้ลองสร้างโมเดลทดลองและทำขึ้นโดยแยกส่วนออกเป็นสามส่วน คือ ส่วนของ knowledge interpreter ซึ่งเป็นส่วนของการเขียนกฎโดยผู้เชี่ยวชาญทางภาษา ส่วนของพจนานุกรม และส่วนของโปรแกรมจัดการระบบความรอบรู้ ปรากฏว่า ใช้แปลภาษาอังกฤษเป็นไทย ในประโยคทดลองได้ดีพอควร

งานประมวลผลภาษาธรรมชาติมีงานให้ทำได้มากมาย

หากพิจารณาโครงการแปลภาษาแล้วจะพบว่าเป็นเรื่องที่จะเข้าสู่จุดสำเร็จได้ยากเพราะจะต้องใช้เวลาอีกมาก แต่จากห้องปฏิบัติการวิจัยไมโครคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ได้ศึกษาวิเคราะห์ภาษาไทยที่ทำให้กลุ่มคณะของผู้เขียนได้ดำเนินงานวิจัยประกอบโครงการทางด้าน NLP ได้หลายโครงการ เช่น โครงการตรวจสอบตัวสะกดภาษาไทย โครงการสร้างพจนานุกรมอิเล็กทรอนิกส์และพจนานุกรมคำพ้อง (thesaurus) เป็นต้น จะเห็นว่าการศึกษา NLP โดยเน้นภาษาไทยเป็นเรื่องที่คนไทยจะต้องดำเนินการเอง เพราะเราทุกคนมั่นใจแล้วว่าไม่มีชนชาติใดในโลกที่จะรู้ภาษาไทยได้ดีเท่ากับพวกเราชาวไทย ดังนั้นพึงภูมิใจว่าเรามีเอกลักษณ์ของชาติที่ทำให้เราเป็นปึกแผ่นและรวมกันเป็นชนชาติได้ก็ด้วยภาษาไทย การประมวลผลภาษาธรรมชาติจึงเป็นศาสตร์ที่พวกเราควรศึกษาและวิจัยกันต่อไป อย่างน้อยก็เพื่อให้ลูกหลานได้รับรู้ต่อไปว่า พวกเราก็ได้พยายามทำกันแล้ว